Project Report

Student Performance in Exam Prediction

Introduction

Student performance data approach student achievement in secondary education. The data attributes include student gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score. We as a group analyzed various parameters and tried to gain more insight into how some of these factors affect the student's performances and tried to predict various metrics relevant to the data set. I have worked on gauging academic performance of students in both subjects. Three performance indicators namely total, average and result. I have made an assumption that total score which is the sum of grades of all the three scores is a performance indicator. In the courses, instructors have to decide the cut-off for each letter grade to be given to students. My performance metrics deals with similar scenario in which we are gauging the student's performance based on comparison of total marks with a certain predefined cut-off. Cut off can vary from one course to another and may vary have different meaning in different scenarios. A lot of competitive exams use percentile as an indicator to define cut - off for performance and use it as an admission criterion. For example, Indian Institute of Management (IIM) conducts one of the most competitive exams in the world in which the performance is gauged based on certain cut-off defined using percentiles. My work is divided 3 segments for each course. First segment involves Exploratory data analysis which is useful for getting familiar with data set and gaining insight into various variables and their interactions. Second segment involves finding significant predictors responsible for determining performance class. This analysis gives us insight into the some of the significant factors which affect student's performance. Third segment involves using prediction models to predict the performance class of the students based on the various features.

Datasets

The training data set is now supplied to machine learning model, on the basis of this data set the model is trained. After the operation of testing, model predict whether the student has been failed in his/her exam or passed in his/her exam.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	total	average	math_pass status	reading_pass status	writing_pass status	result
0	female	group B	bachelor's degree	standard	none	72	72	74	218	72.666667	P	P	P	P
1	female	group C	some college	standard	completed	69	90	88	247	82.333333	P	P	P	P
2	female	group B	master's degree	standard	none	90	95	93	278	92.666667	P	P	P	P
3	male	group A	associate's degree	free/reduced	none	47	57	44	148	49.333333	P	P	P	F
4	male	group C	some college	standard	none	76	78	75	229	76.333333	P	P	P	P
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99	95	282	94.000000	P	P	P	P
996	male	group C	high school	free/reduced	none	62	55	55	172	57.333333	P	P	P	P
997	female	group C	high school	free/reduced	completed	59	71	65	195	65.000000	P	P	P	P
998	female	group D	some college	standard	completed	68	78	77	223	74.333333	P	P	P	P
999	female	group D	some college	free/reduced	none	77	86	86	249	83.000000	P	P	P	P

1000 rows × 14 columns

Steps Involved for Predicting:

- **Understand and define the problem**
- **Analyze and prepare the data**
- **Apply the algorithms:**

  Here in this analysis we have worked out the project with three algorithms to this problem and evaluated its effectiveness. And finally choose the best algorithm and trained the dataset. Here when we worked out with Random Forest Algorithm and KNN Algorithm we got the result as over-fitting. So, we have chosen Logistic Regression Algorithm where when we analyzed the accuracy score was 0.97.

  Logistic Regression**: -** In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc... Each object being detected in the image

would be assigned a probability between 0 and 1 and the sum adding to one. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from log*istic un**it***, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the profit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio. The binary logistic regression model has two levels of the dependent variable: categorical outputs with more than two values are modeled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

- **Predict the result**

```
In [23]:  1  y_pred=classifier.predict(x_test)

In [24]:  1  classifier.predict([[8]])
Out[24]: array(['F'], dtype=object)

In [25]:  1  from sklearn.metrics import accuracy_score

In [26]:  1  score=accuracy_score(y_test,y_pred)

In [27]:  1  score
Out[27]: 0.97
```

Conclusion:

This model helps the organization in making the right decision by predicting whether the student is passed/failed in their respective exams.

.